BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 9 2005, pages 1908–1916
doi:10.1093/bioinformatics/bti315

Structural bioinformatics

Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites

Alasdair T. R. Laurie and Richard M. Jackson*

School of Biochemistry and Microbiology, University of Leeds, Leeds LS2 9JT, UK

Received on December 22, 2004; revised on January 31, 2005; accepted on February 7, 2005

Advance Access publication February 8, 2005

ABSTRACT

Motivation: Identifying the location of ligand binding sites on a protein is of fundamental importance for a range of applications including molecular docking, *de novo* drug design and structural identification and comparison of functional sites. Here, we describe a new method of ligand binding site prediction called Q-SiteFinder. It uses the interaction energy between the protein and a simple van der Waals probe to locate energetically favourable binding sites. Energetically favourable probe sites are clustered according to their spatial proximity and clusters are then ranked according to the sum of interaction energies for sites within each cluster.

Results: There is at least one successful prediction in the top three predicted sites in 90% of proteins tested when using Q-SiteFinder. This success rate is higher than that of a commonly used pocket detection algorithm (Pocket-Finder) which uses geometric criteria. Additionally, Q-SiteFinder is twice as effective as Pocket-Finder in generating predicted sites that map accurately onto ligand coordinates. It also generates predicted sites with the lowest average volumes of the methods examined in this study. Unlike pocket detection, the volumes of the predicted sites appear to show relatively low dependence on protein volume and are similar in volume to the ligands they contain. Restricting the size of the pocket is important for reducing the search space required for docking and *de novo* drug design or site comparison. The method can be applied in structural genomics studies where protein binding sites remain uncharacterized since the 86% success rate for unbound proteins appears to be only slightly lower than that of ligand-bound proteins.

Availability: Both Q-SiteFinder and Pocket-Finder have been made available online at http://www.bioinformatics.leeds.ac.uk/qsitefinder and http://www.bioinformatics.leeds.ac.uk/pocketfinder

Contact: r.m.jackson@leeds.ac.uk

1 INTRODUCTION

The function of a protein is defined by the interactions it makes with other proteins and ligands. Computational methods for the detection and characterization of functional sites on proteins have increasingly become an area of interest (Campbell *et al*., 2003). This is largely due to the many newly solved structures that have poorly characterized biochemical functions or molecular interactions. Faced with a rapidly increasing number of known protein structures, it has become more important to have analytical tools that identify functional sites. This is frequently achieved through functional site detection, which often uses protein evolutionary information (Lichtarge *et al*., 1996; Aloy *et al*., 2001; Armon *et al*., 2001; Landgraf *et al*., 2001) or by structural comparisons of functional sites (Artymiuk *et al*., 1994; Wallace *et al*., 1997; Stark and Russell, 2003). In addition, functional site detection is important for targeting specific sites in structure-based drug design to assist in the development of therapeutic agents. Virtual screening of ligands against protein structures using docking is widely used for identifying potential lead compounds in the drug design process. In addition *de novo* drug design can lead to the creation of novel ligands not found in molecular databases (Honma, 2003). It is essential that the ligand binding site is identified prior to either study as both procedures require this information. Furthermore, all methods can be made more efficient by further restricting the search to critical regions.

Generally, ligand binding site prediction methods analyse the protein surface for pockets. The ligand binding site is usually in the largest pocket. For example, SURFNET (Laskowski, 1995) was used to analyse 67 protein structures (Laskowski *et al*., 1996). The ligand binding site was found to be in the largest pocket in 83% of cases. LIGSITE (Hendlich *et al*., 1997) was used to show that the ligand binding site was found in the largest pocket in all 10 proteins tested. Other pocket detection methods include Cavity Search (Ho and Marshall, 1990), POCKET (Levitt and Banaszak, 1992), CAST (Liang *et al*., 1998), VOIDOO (Kleywegt, 1994), PASS (Brady and Stouten, 2000), LigandFit (Venkatachalam *et al*., 2003), APROPOS (Peters *et al*., 1996) and the methods of Delaney (1992), Del Carpio (1993) and Masuya and Doi (1995). In all cases geometric criteria are used to define the location and extent of the pocket. Q-SiteFinder takes a different approach. The pockets are defined only by energetic criteria. The method calculates the van der Waals interaction energies of a methyl probe with the protein. Probes with favourable interaction energies are retained and clusters of these probes are ranked according to their total interaction energies. The energetically most favourable cluster is then ranked first. It should be noted that there is no requirement that this is also the geometrically largest cluster.

Several techniques have been developed for estimating the interaction energy between a probe at a given point and a protein. One of the most established methods is that developed by Goodford (1985). It identifies sites of favourable interaction with specific probe types. This is particularly useful for structure-based drug design, since it identifies which parts of the protein are likely to interact favourably with functional groups on a drug-like molecule. For example, studies have been carried out to identify the hydrogen bonding potential of drug-like molecules using GRID (Wade and Goodford, 1989; Wade *et al*., 1993). The multiple copy simultaneous search (MCSS) method

*To whom correspondence should be addressed.

of Miranker and Karplus (1991) has also been used to detect favourable binding sites for different functional groups. There are also a number of other methods that have been developed to determine preferential locations for functional groups within binding sites (for a review, see Bohacek and McMartin, 1997). They use interacting probes or fragments with different interaction types such as electrostatic and hydrogen bonding. None of these methods have been used to predict protein–ligand binding sites. However, the methods of Silberstein *et al*. (2003) and Bate and Warwicker (2004) have been applied to enzyme active site detection and that of Ruppert *et al*. (1997) to ligand binding site prediction. The method of Silberstein *et al*. (2003) computationally distributes organic solvent molecules (e.g. acetone, urea, *t*-butanol, etc.) around the surface of an enzyme. The interaction energies between the molecules and the enzyme are optimized using a conventional molecular mechanics function (including van der Waals, electrostatic and solvation terms) in a way similar to the MCSS method. For each type of organic molecule, the distances between the active site and the energy minima were calculated. For six enzymes and six apoenzymes, the enzyme active site was typically identified within 1 Å of one of the five lowest energy minima. Bate and Warwicker (2004) predicted active site location based on the peak of the electrostatic potential. They compared it with the effectiveness of a cleft volume calculation. For 77% of enzymes the electrostatic potential peak is within 5% of that of the surface shell (drawn around the whole molecule) closest to the active site centre as opposed to only 58% of enzymes for which the centre of the largest cleft lies within 5% of the active site. The method of Ruppert *et al*. (1997) has been developed for estimating the interaction energies between a probe at a given point and a protein. Ruppert *et al*. (1997) use the scoring function developed by Jain (1996) to optimize interaction energies of three different probe types (hydrophobic and hydrogen bond donor and acceptor). They retain probes with the most favourable interaction energies. They then identify 'sticky spots', which are regions that have the highest density of probe interaction energy. Next a pocket is grown by defining protein-free spheres in the protein void around the sticky spot. Lastly, a process of accretion takes place, which enlarges the sticky spots into larger pockets, by adding nearby accessible probes defined by the pocket. Thus, both energetic and geometric criteria are used to define a ligand binding site. Their algorithm was shown to give good results on nine ligand-bound proteins and two proteins in the unbound state. In contrast to the above methods, Q-SiteFinder simply uses the van der Waals interaction (of a methyl probe) and an interaction energy threshold to determine favourable binding clefts.

Q-SiteFinder has been designed to meet two main requirements. First, it is intended to be suitable for identification of ligand binding sites for virtual screening and *de novo* drug design. The drug design process requires that the binding site be known as accurately as possible. Second, protein residues within a suitable range of the probe clusters are identified, which could be used for functional site identification and comparison. In both cases it is important to keep the predicted ligand binding site as small as possible without compromising accuracy. In particular, Laskowski *et al*. (1996) demonstrated that pocket size increases linearly with protein volume. This trend is likely to be a geometric property of proteins, as the sizes of ligands are not likely to be related to protein volume. We therefore measure how accurately our predicted sites mapped onto ligand coordinates, and used this measurement to provide a threshold for success. Q-SiteFinder is then compared with a pocket detection algorithm, Pocket-Finder, an implementation of LIGSITE (Hendlich *et al*., 1997). LIGSITE is a widely used pocket detection algorithm. It has been used in defining binding sites in many applications including docking (Rarey *et al*., 1995), *de novo* drug design (Verdonk *et al*., 2001) and in defining binding sites for functional site comparison (Schmitt *et al*., 2002) as well as in the widely used ReliBase, a program for searching a protein–ligand database (Hendlich, 1998).

Table 1. The protein dataset used in this study

1aaq	1bma	1dwd	1hef	1lna	1poc	1tpp	2dbl	3ptb
1abe	1byb	1eap	1hfc	1lpm	1rds	1trk	2gbp	3tpi
1acj	1cbs	1eed	1hri	1lst	1rne	1tyl	2lgs	4aah
1acl	1cbx	1epb	1hsl	1mcr	1rob	1ukz	2mcp	4cts
1acm	1cdg	1eta	1hyt	1mdr	1slt	1ulb	2phh	4dfr
1aco	1cil	1etr	1icn	1mmq	1snc	1wap	2pk4	4est
1aec	1com	1fen	1ida	1mrg	1srj	1xid	2plv	4fab
1aha	1coy	1fkg	1igj	1mrk	1stp	1xie	2r07	4phv
1apt	1cps	1fki	1imb	1mup	1tdb	2ack	2sim	5p2p
1ase	1ctr	1frp	1ive	1nco	1tka	2ada	2yhx	6abp
1atl	1dbb	1ghb	1lah	1nis	1tmn	2ak3	3cla	6rnt
1azm	1dbj	1glp	1lcp	1pbd	1tng	2cgr	3cpa	6rsa
1baf	1did	1glq	1ldm	1pha	1tni	2cht	3gch	7tim
1bbp	1die	1hdc	1lic	1phd	1tnl	2cmd	3hvt	8gch
1blh	1dr1	1hdy	1lmo	1phg	1tph	2ctc	3mth	

2 METHODS

2.1 Datasets and ligand identification

The dataset consisted of 134 records obtained from the Protein Data Bank (PDB) (Berman *et al*., 2000) listed in Table 1. These entries correspond to the GOLD protein–ligand docking dataset described by Nissink *et al*. (2002). All the coordinates in the PDB entries were used. This subset was used instead of all 305 proteins described by Nissink *et al*. (2002) to remove those with high levels of structural similarity (e.g. 1ela, 1elb, 1elc, 1eld and 1ele), which could bias the results.

Coordinates of the ligand(s) were placed in a separate file. Residues covalently bound to the protein were retained in the file containing the protein coordinates. All solvent molecules were discarded (including phosphate, sulphate and metal ions). Q-SiteFinder is not designed to detect the binding sites of small solvent molecules.

We developed a program (LigandSeek) to identify ligand coordinates in PDB files. It divides the atoms in a PDB file into separate groups at TER cards, ATOM/HETATM boundaries or change of chain letter. Groups of non-water atoms that have fewer than 150 atoms are identified as potential ligands. Ligand identification was checked manually by cross-referencing with the Macromolecular Structures Database MSDSite service (Golovin *et al*., 2004) available online at http://www.ebi.ac.uk/msd-srv/msdsite/. LigandSeek was found to be 100% successful in identifying ligands using the 134 proteins described in this study. LigandSeek also identifies residues that the user may not wish to define as a ligand, such as protein phosphotyrosine residues, cofactors such as Haem, peripherally bound carbohydrate residues and small solvent molecules such as SO_4. In the online versions of Q-SiteFinder and Pocket-Finder, options are therefore provided to retain these residues along with protein atoms for binding site analysis or to discard selected residues.

We created a dataset of 35 structurally distinct proteins in the unbound state which share structural similarity with 35 proteins in the ligand-bound dataset. This was achieved through examination of the Structural Classification Of Proteins (SCOP) database (Murzin *et al*., 1995) for the 305 proteins described by Nissink *et al*. (2002). The 305 proteins were used

Table 2. The dataset used in testing Q-SiteFinder with proteins in the unbound conformation

Ligand-bound	Unbound	Ligand-bound	Unbound	Ligand-bound	Unbound
1a6w	1a6u	1mtw	2tga	2h4n	2cba
1acj	1qif	1okm	4ca2	2pk4	1krn
1apu	3app	1pdz	1pdy	2sim	2sil
1blh	1djb	1phd	1phc	2tmn	1l3f
1byb	1bya	1pso	1psn	2ypi	1ypi
1hfc	1cge	1qpe	3lck	3gch	1chg
1icn	1ifb	1rbp	1brq	3mth	6ins
1ida	1hsi	1rne	1bbs	3ptb	2ptn
1igj	1a4j	1snc	1stn	5p2p	3p2p
1imb	1ime	1srf	1pts	6cpa	5cpa
1ivd	1nna	1stp	2rta	6rsa	7rat
1mrg	1ahc	2ctc	2ctb		

rather than just the 134 proteins of the GOLD set to yield enough pairs of homologues. High-resolution structures were favoured where possible. The bound protein–ligand complexes were superimposed onto their unbound homologues. Ligands were then extracted for use with the unbound homologues. Both sets of proteins and ligands were analysed using Q-SiteFinder and the success rates were compared. The protein pairs used in the experiment are shown in Table 2.

2.2 Q-SiteFinder

Q-SiteFinder uses several separate procedures to perform ligand binding site prediction (shown in Supplementary Figure 1). First, ligand coordinates should be separated from the other atom coordinates using LigandSeek. All remaining HETATM records in the protein file are converted to ATOM records, and water molecules removed. Hydrogen atoms are then added to protein atoms by the method described by Jackson *et al.* (1998). The coordinates are rotated about the geometric centre to minimize the volume of the box enclosing the protein. This reduces the number of grid points requiring analysis. The same pre-processing steps are also performed when using Pocket-Finder.

The program Liggrid calculates the non-bonded interaction energy of a probe type with the protein at each position on a defined 3D grid, using the GRID force field parameters as described previously (Jackson, 2002). Here we define the interaction between the protein and a methyl probe ($-CH_3$) at a grid resolution of 0.9 Å on a 3D grid enclosing the whole protein. The probes with the most favourable binding energy are retained based on an interaction energy threshold. A range of values were tested (-1.0 to -1.9 kcal/mol). The probe coordinates are saved in PDB format, and the coordinates are rotated back to match the original orientation of the protein. Individual probe coordinates are then clustered according to their spatial proximity, and the total interaction energies of probes within each cluster are calculated. Probe clustering uses a variable known as the connection range, which determines the maximum distance between two probes that can be connected as part of the same cluster. This value should be greater than the probe grid resolution used to generate the probe output file. The default used here is 1.0 Å for a grid resolution of 0.9 Å. This connects all adjacent sites but not those on the diagonals of the cube. The probe clusters are ranked according to their total interaction energies, with the most favourable being identified as the first predicted binding site. The speed of the overall process is dependent on protein size, but it is usually 10–15 s on the current server (1.8 GHz CPU).

The Clustering program also calculates site volume, and can identify which protein atoms are within a defined range of cluster sites. It is also used in this capacity in Pocket-Finder (discussed below). The parameters for estimation of site volume and identification of protein residues are different for Q-SiteFinder and Pocket-Finder. Values of 5.0 and 3.0 Å are used, respectively, to identify protein atoms in contact with the site. For the volume calculation, a distance threshold was used to calculate the number of cubes of dimension 0.5 $Å^3$ within 2.0 and 1.0 Å, respectively, of the probe sites. These values reflect the fact that probe sites identified with Q-SiteFinder approach the protein within van der Waals (vdW) contact, i.e. the sum of the two vdW radii, as opposed to Pocket-Finder where sites approach the vdW surface of the protein, i.e. the vdW radius of protein atoms. This was found to produce sites in both cases with approximately a single layer of protein atoms surrounding the probes and approximately the same site volume.

2.3 Pocket-Finder

Pocket-Finder implements LIGSITE (Hendlich *et al*., 1997) which is based on the POCKET algorithm (Levitt and Banaszak, 1992). In POCKET, a probe sphere of radius 3 Å is passed across the protein along each line of a 3D grid in the x, y and z directions. An interaction between the protein and probe sphere occurs if the centre of a protein atom is found inside the probe sphere. A pocket is identified if an interaction occurs followed by a period of no interaction, followed by another interaction. This is referred to as a protein–site–protein (PSP) event. The definition of the pocket is somewhat dependent on the angle of rotation of the protein relative to the axes. LIGSITE improves on POCKET by scanning along the four cubic diagonals in addition to the x, y and z directions. This makes the identification of protein pockets much less dependent on the orientation of the protein on the 3D grid. Like LIGSITE, Pocket-Finder measures the extent to which each grid point is buried in the protein. Each grid point has seven scanning lines passing through it (in the x, y and z directions and the four cubic diagonals). The grid points are initially set to zero. Every time a grid point is identified as being in a pocket in a PSP event, the grid point is incremented by one. Grid points can therefore register from zero (not part of a pocket) to seven (deeply buried in a cavity) PSP events. Grid points are only retained if they exceed a threshold number of PSP events. The threshold is termed the MINimum PSP (MINPSP). Pockets are defined by cubes of retained grid points with sides of length equal to the grid resolution. We use a grid resolution of 0.9 Å, a probe radius of 1.6 Å and a MINPSP of 5. These values reduce the average volume of the first predicted site when compared with the parameters used by Hendlich *et al*. (1997) (grid resolution of 0.5 or 0.75 Å, a probe radius of 1.4 Å and a MINPSP of 2).

Pocket-Finder generates a probe output file that is compatible with the clustering method (described above). However, the sites produced by the Pocket-Finder program are ranked according to the number of probes in the site rather than by probe energy.

2.4 Protein volume calculation

PDBVolume gives an estimate of the protein volume. It is a requirement that the PDB file is first pre-processed (described above). PDBVolume creates a 3D grid with resolution 0.9 Å and places a probe (radius 1.7 Å) at each position. If the probe overlaps with a protein atom, the grid point is marked as being occupied. The number of cubes with sides of length 0.9 Å and a marked grid point at each vertex are counted to estimate the volume. A comparison between protein volume calculations carried out by Laskowski *et al*. (1996) (for the 12 proteins labelled in their graphs) and PDBVolume gave a standard deviation of 3.3%. PDBVolume was also used to calculate ligand volume. Hydrogen atoms were added to the ligands and a higher grid resolution of 0.1 Å (rather than 0.9 Å) was used to calculate volumes.

3 RESULTS AND DISCUSSION

3.1 Comparison of Q-SiteFinder and Pocket-Finder

Q-SiteFinder analyses clusters of energetically favourable methyl binding sites to predict the ligand binding sites. Three sets of results are presented here: development and calibration of the method,

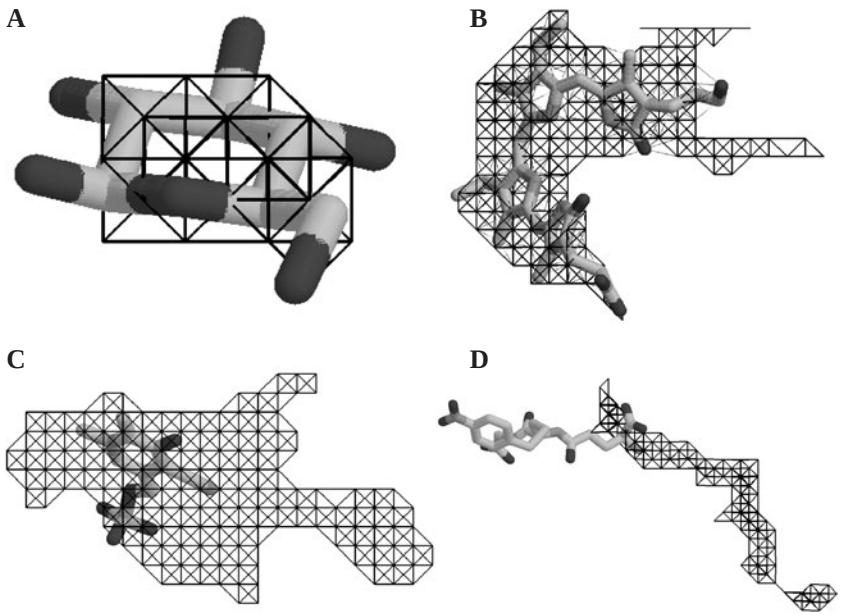

Fig. 1. Examples of different levels of predicted binding site precision (for a definition of precision, see text). (**A**) 2gbp, 100% (Q-SiteFinder); (**B**) 1bbp, 68% (Q-SiteFinder); (**C**) 1asc, 26% (Pocket-Finder); (**D**) 1glq, 17% (Q-SiteFinder). Probe centres are shown in black wireframe.

comparison with two pocket-detection algorithms and testing its ability to predict ligand binding sites on proteins in the unbound state.

We measure how well a predicted site maps onto the ligand coordinates using a precision threshold. The term 'precision' used here defines the percentage of probe sites in a single cluster that are within 1.6 Å of a ligand atom. We define a successful prediction using a precision threshold.

A threshold of 25% precision was used to define success in all the results presented here. For example, the predicted site shown in Figure 1C with a precision of 26% is considered a success; however, the site shown in 1D with a precision of 17% is not. We feel this is a very stringent measure of success, since (1) a significant number of probe sites must be within the range of the ligand and (2) simply predicting very large pockets that include the ligand binding sites will not be counted as a success. It should be noted that a method that includes the entire protein surface in a single 'pocket' will be 100% successful unless such a precision threshold is used. However, such a prediction is of little utility for guiding docking studies, *de novo* drug design or functional site comparisons.

In addition to a 25% precision threshold, the following conditions apply to the definition of success. If a ligand is successfully predicted in more than one site on a protein, it is counted as a success only in the higher ranking site, since these predicted sites can be considered to be part of the same binding site. If more than one ligand is found in the same site, only the success with the highest precision is counted for this site. This affected only four cases: 1glp, 1glq, 1ukz and 2phh. If two ligands are successfully predicted in two different sites on a protein, these are counted separately.

The results have been derived using the coordinates of 134 structures corresponding to the GOLD docking test set described by Nissink *et al*. (2002). Their actual coordinates were not used, since they contain only the binding site and surrounding atoms. The coordinates were taken in their entirety from the PDB entries (Table 1) using all protein chains and not solely single subunits.

Figure 2A shows the results of using Q-SiteFinder with a range of energy threshold values (-1.0 to -1.9 kcal/mol) for retaining methyl binding sites. The maximum success rate was achieved when a binding energy cut-off of -1.4 kcal/mol was used. This cut-off was used to generate the other results presented in this report. The success rate was 71% in the first predicted binding site, and the average precision was 68%. It is desirable to have both a high rate of success and a high precision of binding site prediction. Figure 1B shows an example of 68% precision, giving an idea as to the average capabilities of Q-SiteFinder. The average volume of the first predicted site was 390 $Å^3$ (1% of the average protein volume). However, this varies between 0.2 and 3.0% of the protein volume. There was at least one successful prediction in the top three predicted sites for 90% of the proteins, and at least one successful prediction in the top ten predicted sites for 96% of the proteins.

Pocket-Finder uses a variable, MINPSP, the minimum number of PSP events (see Methods). This can be thought of as a burial threshold, and PSP values for each grid point vary from 0 (not a pocket) to 7 (deeply buried). Figure 2B shows that the best success rate for Pocket-Finder is obtained at a MINPSP threshold of 5. The success rate is only 48% in the first predicted site with an average precision of 29%. There was at least one successful prediction in the top three predicted sites for 65% of the proteins, and at least one successful prediction in the top ten predicted sites for 74% of the proteins. The average volume of the first predicted site is 1300 $Å^3$ (3% of the average protein volume). Hendlich *et al*. (1997) recommend a MINPSP of 2. In our implementation of Pocket-Finder this gives a relatively low average precision (8%) and a relatively large site volume of 8700 $Å^3$ (23% of the average protein volume). No significant benefit in the success rate was observed on using a MINPSP of 2 rather than 5 when the minimum threshold for success (more than

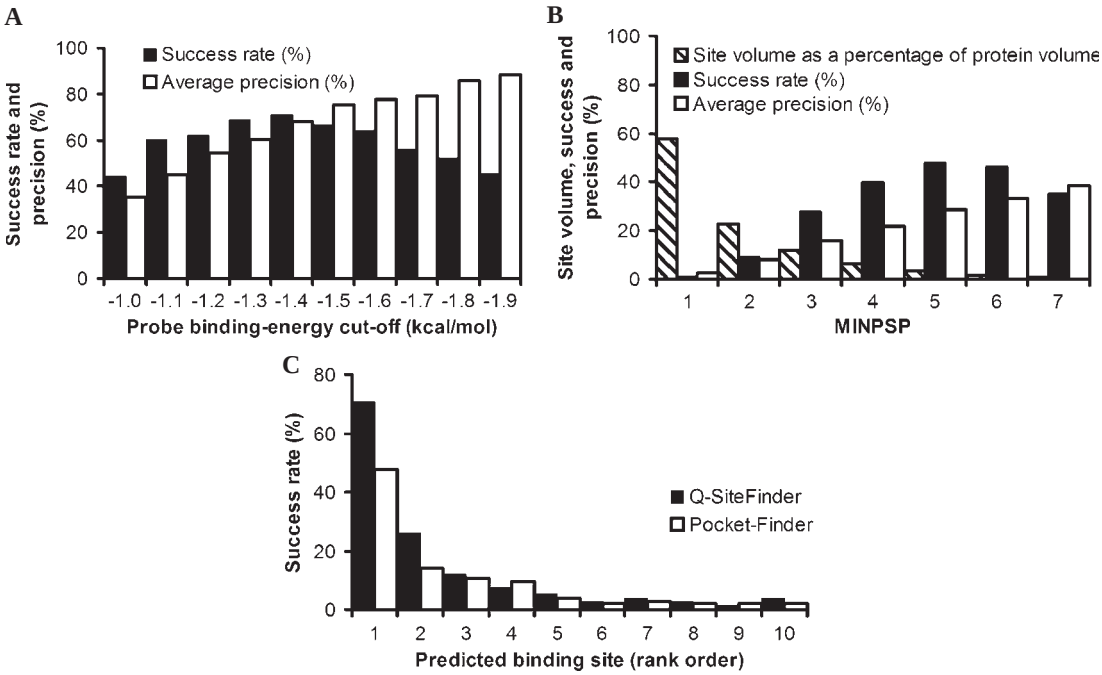

Fig. 2. (**A**) The success rates (in the first predicted binding site) and the average precision when different probe binding-energy cut-offs are used in Q-SiteFinder. Complete failures (i.e. a precision of 0%) were excluded from the calculation of the average precision values. (**B**) The average volumes, success rates and the average precisions for the first predicted site when different MINPSP thresholds (see Methods) are used in Pocket-Finder. (**C**) A comparison of the success rates for Q-SiteFinder and Pocket-Finder for the top ten predicted sites (a probe binding-energy cut-off of −1.4 kcal/mol was used for Q-SiteFinder and a MINPSP threshold of 5 was used for Pocket-Finder).

0% precision) was used. A MINPSP value of 5 was used to generate the other results presented in this report. Figure 2B also shows the relationship between site volume and precision. Smaller sites have a higher average precision. This is expected, since sites with high volumes will usually incorporate locations on the protein surface that are not part of the binding site. It is interesting to note that a MINPSP of 7 still gives a relatively high success rate. Such grid points form part of a cavity, since they are bound on all sides by protein. This suggests that about one-third of the proteins in our dataset undergo a conformational change on binding that completely encloses the ligand. A comparison between the success rates for Q-SiteFinder and Pocket-Finder is shown in Figure 2C. Q-SiteFinder has a higher success rate in each of the top three predicted binding sites.

Q-SiteFinder did not identify the ligand binding site for 1cdg, 1eta, 3cla and 2yhx (which had 0% precision for all predicted sites) and 1glq and 1glp (which had partial successes in the first predicted site, below the 25% precision threshold). 1cdg represents the structure of cyclodextrin glycosyltransferase. It has three maltose sugar moieties which bind at the protein surface, and are in very shallow clefts. Large probe clusters are therefore not generated at these sites. However, the catalytic site of the protein is in a cleft, and binds to cyclodextrin (Uitdehaag *et al*., 1999) in other ligand-complexed PDB entries for this enzyme. The fourth predicted site identifies this binding site and is within 5.0 Å of contacting residues Arg-47 and Asp-371. This success was not identified during analysis because the coordinates of cyclodextrin are not present in the 1cdg structure. 1eta is a thyroxine transporter (Hamilton *et al*., 1993). However, only one symmetrical unit (a dimer) is described by the PDB coordinates used in this study. The biologically relevant tetramer forms two thyroxine binding sites

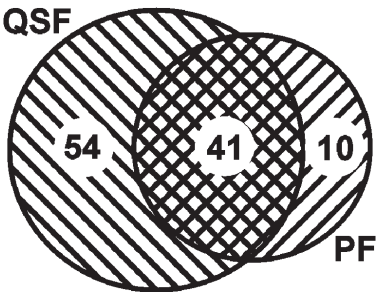

Fig. 3. Overlap in ligand binding site prediction in the first predicted site. Pocket-Finder (PF) predicts 10 sites that were not predicted by Q-SiteFinder (QSF). Q-SiteFinder predicts 54 sites that were not predicted by Pocket-Finder and 41 sites are predicted by both methods.

between two symmetrical units. When analysis was performed on the tetramer [coordinates taken from the PQS database (Henrick and Thornton, 1998)], the two binding sites were successfully identified by Q-SiteFinder in the first and third predicted sites. Similarly, 3cla is a trimer formed from three symmetrical units. Only a single unit was described by the PDB coordinates. When the trimer was analysed with Q-SiteFinder, the three ligand binding sites were identified in the top three predicted sites (albeit with precisions below the 25% threshold).

There was a fairly high degree of overlap in the detection of ligand binding sites by Q-SiteFinder and Pocket-Finder (Fig. 3). Pocket-Finder identified only 10 ligand binding sites that were not identified by Q-SiteFinder in the first predicted site. However, all 10 were

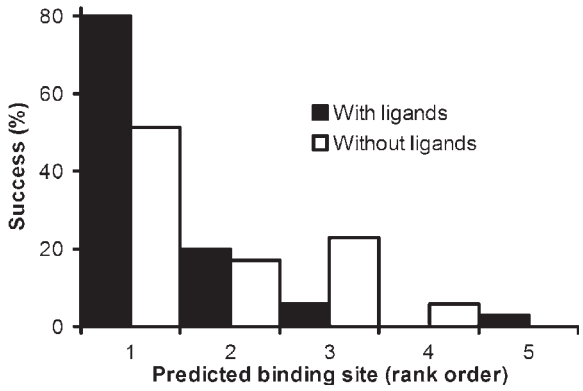

Fig. 4. Success rates of binding site prediction when Q-SiteFinder was used for 35 ligand-bound proteins and 35 unbound homologues.

identified by Q-SiteFinder in the second or third predicted sites. Q-SiteFinder identified 54 that were not identified by Pocket-Finder. Therefore, Pocket-Finder detects a subset of the ligand binding sites detected by Q-SiteFinder.

3.2 Application of Q-SiteFinder for detecting binding sites on unbound proteins

It is anticipated that Q-SiteFinder will be used to detect binding sites on proteins that are not bound to ligands. It is possible that ligand binding may cause a conformational change in the protein that biases the program to select a particular site. To test unbound conformations, 35 structurally distinct unbound proteins were compared with 35 homologous ligand-bound proteins as described in the Methods section.

Figure 4 shows that the success rate in the first predicted site was lower for the unbound state (51%) than for the ligand-bound state (80%). The percentages of proteins with at least one success in the top three sites were 86% for the unbound state and 97% for the ligand-bound state. The average precision of the first predicted binding site (excluding total failures) was 71% for the unbound state and 74% for the ligand-bound state.

The reduced success rate for the unbound conformation is caused by a number of factors. In two cases (1acj/1qif and 1snc/1stn), subtle changes in the protein structures meant that the predicted sites in the unbound form fell below the 25% precision threshold for success. In some cases, the structure of the ligand binding site was significantly different in the unbound conformation; for example, 1byb/1bya and 1ida/1hsi. 1byb and 1bya are structures of β-amylase (Mikami *et al*., 1994). In the ligand-bound conformation (1byb), the VAL-99–GLY-100–ASP-101 loop appears to fold over the maltotetraose ligand. However, in the unbound conformation, the loop folds away from the binding site. This alters the structure of the binding site, but it is still successfully identified by Q-SiteFinder in the fourth predicted site compared with the first predicted site in the bound conformation (Fig. 5A). 1ida (Tong *et al*., 1995) and 1hsi (Chen *et al*., 1994) are structures of the HIV protease; the unbound form undergoes a sizable induced fit on ligand binding. The main chain of the ligand binding site of the unbound form (1hsi) is much more open. This reduces the interaction in the binding site and, consequently, no large probe clusters are formed (Fig. 5B).

3.3 Predicted site volume

Figure 6A and B show the relationship between the predicted cleft volume of the first predicted binding site and the protein volume for Q-SiteFinder and Pocket-Finder. The results can also be compared with those of SURFNET (Laskowski *et al*., 1996).

Q-SiteFinder produced the smallest first predicted binding sites of 390 Å^3 on average, which shows the best agreement with the average ligand volume (275 Å^3). The average volumes of the first predicted sites as a percentage of protein volume were 15% (SURFNET class 1), 8% (SURFNET classes 2 and 3), 3% (Pocket-Finder) and 1% (Q-SiteFinder). The volumes of the sites predicted by Q-SiteFinder are only weakly dependent on protein volume (Fig. 6B). No predicted site exceeds 1200 Å^3 even at very large protein volumes. This trend closely parallels the relationship between protein volume and the volume occupied by the ligand where there is little correlation between protein volume and ligand volume (Fig. 6C). However, for the pocket detection algorithms, the size of the pocket is more closely related to protein volume; therefore, as protein volume increases, so does the average volume of the first predicted pocket. Hence, Q-SiteFinder predicts sites with volumes that are most appropriate for the size definition of a ligand binding site. Figure 6A shows that SURFNET produced the largest first predicted binding sites on average. However, SURFNET has the highest success rate (83.6%) of all the methods in the first predicted site.

No significant difference was noted between the volumes of successful predictions and unsuccessful predictions for Q-SiteFinder in the first predicted site. Interestingly, for Pocket-Finder, the average volume of successful predictions in the first predicted site was 460 Å^3, much less than the average volume of unsuccessful sites (2100 Å^3). This is because the precision threshold of 25% ensures that predictions defined as a success map well onto the ligand coordinates. Bigger sites often encompass large areas that are not occupied by ligand atoms.

3.4 Precision and comparison with previous studies

The threshold for success used in this study requires that at least 25% of the probe sites in a single cluster are within 1.6 Å of a ligand atom. This implies a precision value greater than 25%. In previous studies no precision threshold has been applied, the only criterion being that the ligand is found somewhere in the predicted pocket. If we relax our threshold to allow any non-zero value (success requires a precision greater than 0%) then at least one ligand atom must be situated in a predicted pocket of infinite size. In this case, the success rates of Pocket-Finder approach those of Q-SiteFinder (Fig. 7A). However, this is at the cost of a significant increase in the volume of the cavity for Pocket-Finder (Fig. 7B), both in comparison with Pocket-Finder at a precision greater than 25% and relative to Q-SiteFinder at a precision greater than 0%. Indeed, there is little change both in the success rate or the average volume of predicted sites for Q-SiteFinder in going from a precision threshold of 0–25%. This implies that the method is relatively insensitive to change in the precision threshold unlike Pocket-Finder. This is due to the fact that the average precision of Pocket-Finder is 29% while that of Q-SiteFinder is 68%. Hence, Q-SiteFinder would appear to be more robust than Pocket-Finder, and better able to pinpoint the location of the ligand binding site.

Furthermore, there is little difference (2%) between the success rate for Pocket-Finder with a MINPSP of 2 and that with 5 despite a 4-fold reduction in the average predicted site volume between these

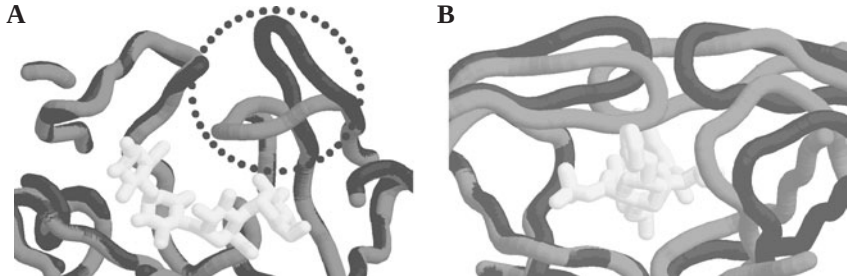

Fig. 5. Backbone structures of homologous ligand-bound (mid-grey) and unbound (dark grey) proteins have been superimposed with their ligands (light grey). (**A**) 1byb (mid-grey) and 1bya (dark grey). The VAL-99–GLY-100–ASP-101 loop has been circled. (**B**) 1ida (mid-grey) and 1hsi (dark grey).

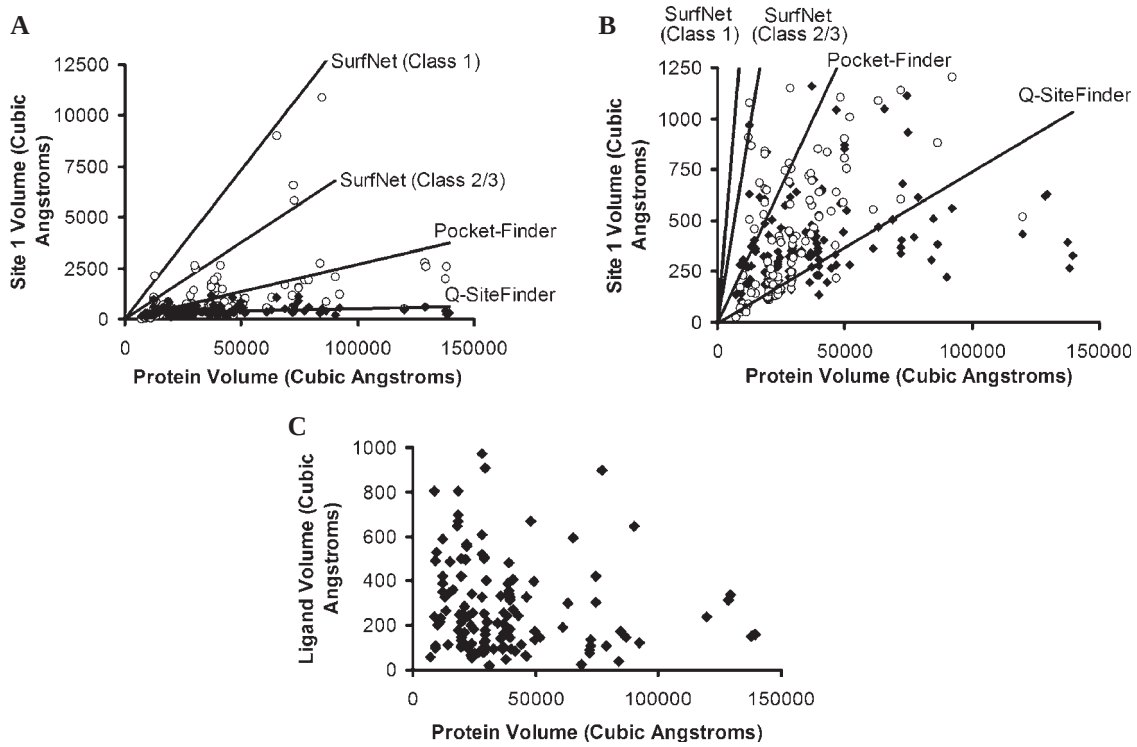

Fig. 6. Volume of the first predicted site as a function of protein volume for Q-SiteFinder, Pocket-Finder and SURFNET. Class 1 enzymes are defined by Laskowski *et al*. (1996) to have the ligand in the first predicted site. Class 2 enzymes are defined to have the ligand in the second predicted site. The remaining proteins are defined as Class 3. (**A**) Volume of the first predicted site as a function of protein volume for Q-SiteFinder, Pocket-Finder and SURFNET. Data for individual proteins are marked as white circles (Pocket-Finder) and black diamonds (Q-SiteFinder). Only the lines of best-fit are shown for SURFNET. Data for 1wap are not shown here due to the unusually large predicted binding site generated by Pocket-Finder (44 000 Å^3). (**B**) The same graph is shown again with a different *y*-scale. All of the Q-SiteFinder data are shown, but only part of the Pocket-Finder data are shown. (**C**) Ligand volume as a function of protein volume for all ligands in the 134 protein dataset. *Note:* Following Laskowski *et al*. (1996), the lines of best-fit ($y = mx$) pass through the origin.

values (data not shown). Therefore, increasing the pocket size (by decreasing MINPSP) does not significantly increase the success rate of Pocket-Finder. It can be concluded that ligands have a preference for regions of the protein that are more buried (Pocket-Finder) and better able to participate in van der Waals interactions with the protein (Q-SiteFinder).

Precision is a useful method for measuring how well probes map onto ligand coordinates (Fig. 8). The main disadvantage of precision is that a high score can be achieved if the probe cluster maps accurately onto only a part of the ligand. In many cases, this is justified, since only a part of the ligand may be bound to the protein. However, in some cases, a high precision can be achieved even though a part of the ligand bound to the protein has not been identified by the probe cluster.

Other studies have used different measures of success. For example, Peters *et al*. (1996) defined a successful prediction as one that includes at least seven of the protein atoms in contact with the ligand. This definition of success has two major problems. First, a very large predicted site (such as one that spreads across the whole surface of the protein) would be considered successful providing

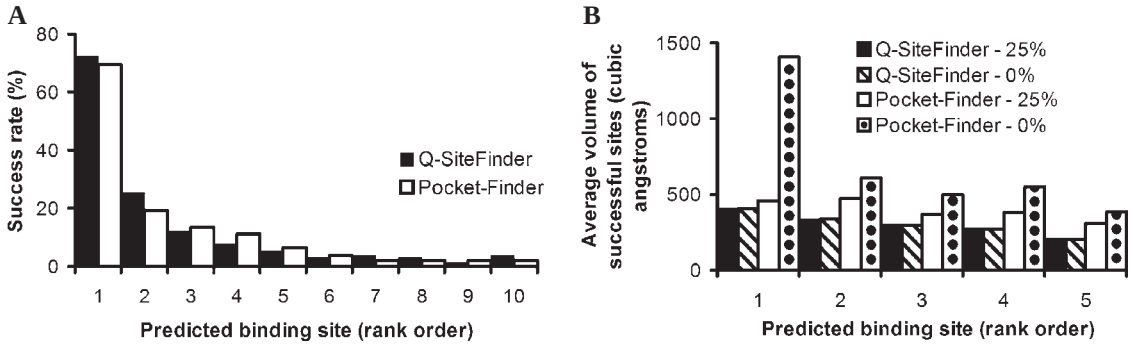

Fig. 7. Success rates of Q-SiteFinder and Pocket-Finder when the threshold for success requires a precision greater than 0%. (**A**) A comparison between Q-SiteFinder and Pocket-Finder for the top 10 predicted sites. (**B**) Average volumes of successfully predicted sites, when 0 and 25% precision thresholds are used to define success in Pocket-Finder and Q-SiteFinder.

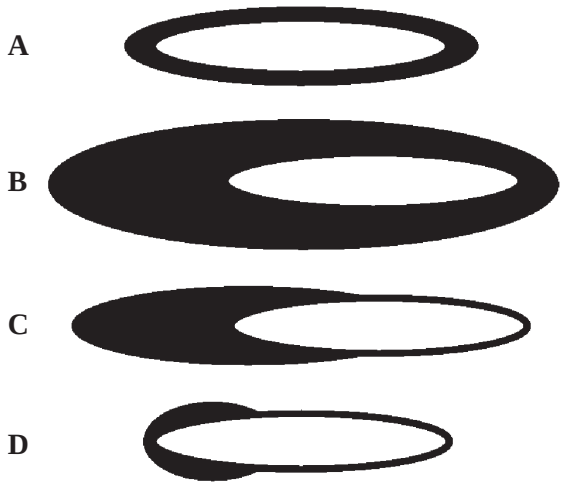

Fig. 8. Different levels of precision. The ligand is shown in white and the probe cluster is shown in black. (**A**) High precision with all of the ligand covered. (**B**) Low precision with all of the ligand covered. (**C**) Low precision with part of the ligand covered. (**D**) High precision with part of the ligand covered.

it incorporated at least seven protein atoms in contact with ligand atoms, even though such a site would be very imprecise. False positive protein residues are not taken into account. Second, if fewer than seven protein atoms were in contact with the ligand, no prediction could be defined as a success even if all of the protein atoms in contact with the ligand were correctly identified.

Hendlich *et al*. (1997) measured the accuracy of their LIGSITE algorithm by finding the percentage of protein atoms that formed part of a pocket that were in contact with ligand atoms. Protein and ligand atoms were defined to be in contact with each other if they were within a distance of the sum of the van der Waals radii plus 0.5 Å. They used a test set of 10 proteins and found that 100% of the contacting atoms were identified in each case. The main disadvantage of this method is that false positive protein residues are not taken into account. If the entire surface of a protein were identified as a predicted binding site, it would score 100%.

Ruppert *et al*. (1997) used three different probe types (hydrophobic and hydrogen bond donor and acceptor probes). They measured the success of their predictions by finding the maximum, minimum and average distances between ligand atoms and the nearest probe whose type matched the ligand atom in question. The reported distances were low. However, this method for calculating success disregards all probes that bind further away from the ligand (false positives). Hence good results could be reported even if the predicted site was very large (for example, covering the entire surface of the protein).

'Precision' is a way of measuring the extent to which a predicted site maps onto ligand coordinates. A method that gives a high precision is a suitable starting point for ligand docking studies, *de novo* drug design and functional site definition. Hence, we conclude that a precision-based threshold for success is suited to measuring the ability of a method to achieve this aim.

4 CONCLUSIONS

We have presented a method, Q-SiteFinder, for ligand binding site prediction that is based on determining energetically favourable binding sites on the surface of a protein. The method is better able to pinpoint the location of the ligand binding site than a comparable pocket detection algorithm (Pocket-Finder) on a dataset of 134 proteins. One of the strengths of the method is its prediction of relatively small sites. The sites have volumes roughly equivalent to ligand volumes irrespective of the overall size of the protein. This is in contrast to pocket detection, where predicted site volumes show a much greater tendency to increase with protein size. This property would appear to be a result of using probe site binding energies with the appropriate energy cut-off rather than purely geometric criteria to determine favourable binding sites on proteins. The individual probe sites relate most closely to the favoured high-affinity binding sites on the protein surface. These favourable binding sites relate to locations where a putative ligand could bind and optimize its van der Waals interaction energy. Such sites would be expected to correspond closely to a high-affinity ligand binding site. This is supported by the high level of success of the method. First, it would appear that this measure is general enough to be of predictive value for a broad range of proteins and ligands of different chemical composition. Furthermore, given the high level of success in unbound protein sites, it is also a property of binding sites that do not have a ligand already bound.

Q-SiteFinder was shown to identify sites with high precision. The advantage of this is that putative binding sites are identified

as closely as possible to the actual binding site. It is important to keep the predicted ligand binding site as small as possible without compromising accuracy for a range of applications such as molecular docking, *de novo* drug design and structural identification and comparison of functional sites.

ACKNOWLEDGEMENT

We would like to thank the BBSRC for financial support in the form of a studentship to A.T.R.L.

REFERENCES

Aloy,P. *et al*. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.

Armon,A. *et al*. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, **307**, 447–463.

Artymiuk,P.J. *et al*. (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.*, **243**, 327–344.

Bate,P. and Warwicker,J. (2004) Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.*, **340**, 263–276.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bohacek,R.S. and McMartin,C. (1997) Modern computational chemistry and drug discovery: structure generating programs. *Curr. Opin. Chem. Biol.*, **1**, 157–161.

Brady,G.P.,Jr and Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, **14**, 383–401.

Campbell,S.J. *et al*. (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.

Chen,Z. *et al*. (1994) Crystal structure at 1.9-Å resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. *J. Biol. Chem.*, **269**, 26344–26348.

Del Carpio,C.A. *et al*. (1993) A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions. *J. Mol. Graph*, **11**, 23–29.

Delaney,J.S. (1992) Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph*, **10**, 174–177.

Golovin,A. *et al*. (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **32** (Database issue), D211–D216.

Goodford,P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857.

Hamilton,J.A. *et al*. (1993) The x-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val-30–>Met variant to 1.7-Å resolution. *J. Biol. Chem.*, **268**, 2416–2424.

Hendlich,M. (1998) Databases for protein–ligand complexes. *Acta Crystallogr. D*, **54**, 1178–1182.

Hendlich,M. *et al*. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model*, **15**, 359–363.

Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.

Ho,C.M. and Marshall,G.R. (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput. Aided Mol. Des.*, **4**, 337–354.

Honma,T. (2003) Recent advances in *de novo* design strategy for practical lead identification. *Med. Res. Rev.*, **23**, 606–632.

Jackson,R.M. (2002) Q-fit: a probabilistic method for docking molecular fragments by sampling low energy conformational space. *J. Comput. Aided Mol. Des.*, **16**, 43–57.

Jackson,R.M. *et al*. (1998) Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, **276**, 265–285.

Jain,A.N. (1996) Scoring noncovalent protein–ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided Mol. Des.*, **10**, 427–440.

Kleywegt,G.J. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D*, **50**, 178–185.

Landgraf,R. *et al*. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, **307**, 1487–1502.

Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*, **13**, 323–330.

Laskowski,R.A. *et al*. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.

Levitt,D.G. and Banaszak,L.J. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph*, **10**, 229–234.

Liang,J. *et al*. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.

Lichtarge,O. *et al*. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Masuya,M. and Doi,J. (1995) Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *J. Mol. Graph*, **13**, 331–336.

Mikami,B. *et al*. (1994) Crystal structures of soybean beta-amylase reacted with beta-maltose and maltal: active site components and their apparent roles in catalysis. *Biochemistry*, **33**, 7779–7787.

Miranker,A. and Karplus,M. (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins*, **11**, 29–34.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nissink,J.W. *et al*. (2002) A new test set for validating predictions of protein–ligand interaction. *Proteins*, **49**, 457–471.

Peters,K.P. *et al*. (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.*, **256**, 201–213.

Rarey,M. *et al*. (1995) Time-efficient docking of flexible ligands into active sites of proteins. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 300–308.

Ruppert,J. *et al*. (1997) Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.*, **6**, 524–533.

Schmitt,S. *et al*. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387–406.

Silberstein,M. *et al*. (2003) Identification of substrate binding sites in enzymes by computational solvent mapping. *J. Mol. Biol.*, **332**, 1095–1113.

Stark,A. and Russell,R.B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res.*, **31**, 3341–3344.

Tong,L. *et al*. (1995) Crystal structures of HIV-2 protease in complex with inhibitors containing the hydroxyethylamine dipeptide isostere. *Structure*, **3**, 33–40.

Uitdehaag,J.C. *et al*. (1999) The cyclization mechanism of cyclodextrin glycosyltransferase (CGTase) as revealed by a gamma-cyclodextrin–CGTase complex at 1.8-Å resolution. *J. Biol. Chem.*, **274**, 34868–34876.

Venkatachalam,C.M. *et al*. (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph Model*, **21**, 289–307.

Verdonk,M.L. *et al*. (2001) SuperStar: improved knowledge-based interaction fields for protein binding sites. *J. Mol. Biol.*, **307**, 841–859.

Wade,R.C. and Goodford,P.J. (1989) The role of hydrogen-bonds in drug binding. *Prog. Clin. Biol. Res.*, **289**, 433–444.

Wade,R.C. *et al*. (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.*, **36**, 140–147.

Wallace,A.C. *et al*. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.